一种支持双栈及高速网络的数字资源利用分析系统数据过滤方法

潘竹虹 [1]　萧德洪 [1,2]
[1](厦门大学信息与网络中心　厦门 361005)
[2](厦门大学图书馆　厦门 361005)

摘要: 【目的】促进高校图书馆数字资源的合理建设、科学管理和高效利用。【应用背景】IPv6 的推广和万兆校园网的普及给网络数据采集造成困难。【方法】提出一种网络设备端口镜像设计方法，在数字资源利用分析系统采集数据前进行 IPv4 及 IPv6 网络数据过滤。【结果】实际部署一个支持 IPv4/IPv6 双栈及万兆网络的数字资源利用分析系统。【结论】使图书馆数字资源利用分析系统适应双栈高速的校园网网络环境。

关键词: 数字资源　信息采集　IPv6　端口镜像
分类号: TP393.1　G25

1 引　言

高校图书馆数字资源不仅包括日常使用的文献资料，还包含图书馆网络平台中共享的信息、软件、国内外数据库等网络学术资源。如何促进高校图书馆数字资源的合理建设、科学管理和高效利用，是当前图书馆工作关注探究的问题之一，并已有较多关于图书馆数字资源利用评估体系的研究[1]及实践探索。

数字资源利用评估常见调查手段为用户调查问卷报告[2]以及数据库供应商提供的访问记录等，其中数字资源登录检索及下载的次数、用户群与用户比例等经由网络传递的数据信息可由校园网网络环境获取，据了解目前已有部分高校图书馆部署了自主研发或者商用的数字资源利用分析系统。受限于专用硬件高昂的成本，常见的数字资源利用分析系统通常以通用处理器实现包头软解析，此类系统的网络采集瓶颈为 1Gbps[3]，并且一般未支持 IPv6。随着 IPv6 在高校的推广以及万兆校园网的普及，数字资源利用分析系统已经难以承载校园网流量。

本文致力于设计一种支持 IPv4/IPv6 双栈以及万兆网络环境的数据过滤方法，该方法在系统采集数据之前完成数据过滤，仅将与图书馆相关的 IPv4/IPv6 流量输出至数据采集与分析功能模块，使数字资源利用分析系统能适应高速双栈的高校校园网环境。

2 需求及设计思路

2.1 技术背景

(1) IPv6 的推广

IPv6 是 IETF 设计的用于替代现行版本(IPv4)的下一代 IP 协议，于 1994 年在 RFC1752 中被定义[4]。由国家发展和改革委员会主导的下一代互联网示范工程(CNGI)项目已于 2010 年完成百所高校的 IPv6 全面覆盖[5]。2012 年 6 月 6 日，全球范围内的 IPv6 网络正式启动[6]。IPv6 网络已日渐推广并承载了部分数字资源传输，但由于协议体系结构的不同，网络管理方面也存在很大差异，面向 IPv6 的管理分析工具不完善，IPv6 成为图书馆数字资源利用分析系统的难题之一。

(2) 万兆校园网的普及

2002 年 7 月 18 日 IEEE 通过了 802.3ae 万兆以太网标准后，万兆以太网凭借其高达 10Gbps 的带宽以

通讯作者：潘竹虹，ORCID：0000-0002-6837-5349，E-mail：zhpan@xmu.edu.cn。

及种种技术优势，逐渐在高校校园网中普及应用，同时也给高校图书馆数字资源利用分析系统带来了极大压力。

网络数据采集可分为基于专用硬件的网络数据采集和基于通用处理器平台的网络数据采集。基于专用硬件的采集方法在高速链路的环境下有巨大的性能优势，但是昂贵的成本让大多数系统仍选用基于通用处理器以软件方法实现包头解析，受限于操作系统和硬件性能，目前处理速度仅能达到 1Gbps(1488Kpps)[3]，万兆流量链路已成为数字资源利用分析系统部署的瓶颈。

(3) 流镜像技术

在数据捕获采集前，由网络设备进行数据输出时的过滤分流是解决系统网络采集瓶颈的有效手段。

常见的数据输出技术为端口镜像及分路器。分路器为物理层分接设备，不支持数据过滤。端口镜像技术是网络设备提供的管理功能，将指定端口或 VLAN 的报文复制一份到其他端口[7]。流镜像技术(基于 ACL 的镜像)是端口镜像技术的一种，仅将匹配流分类条件的报文复制到指定目的地[8]，是目前唯一具有数据过滤功能的镜像技术。

端口镜像技术需要消耗交换机的软硬件性能，当前国内外顶级交换机的镜像输出能力也局限于 4 组甚至少于 2 组镜像组。流镜像技术仅在部分最新网络设备的最新软硬件版本中被支持，功能有较多局限，稳定性也需验证，一般仅支持一个输出[8]。流镜像技术可支持数字资源利用分析系统的数据过滤需求，然而大多数校园网核心设备并不支持流镜像技术，支持该技术的设备上有限的流镜像资源也往往已用于支持用户分析、攻击防护、安全检测、舆情控制等网络安全管理系统。

综上所述，流镜像技术局限性大、部署成本高且仅支持少量输出，难以在校园网核心设备上部署该技术为系统提供数据过滤。

2.2　复合镜像技术

端口镜像技术可细分为三种：本地镜像、远程镜像、流镜像。图 1 为各种镜像技术的镜像流量走向示意图。

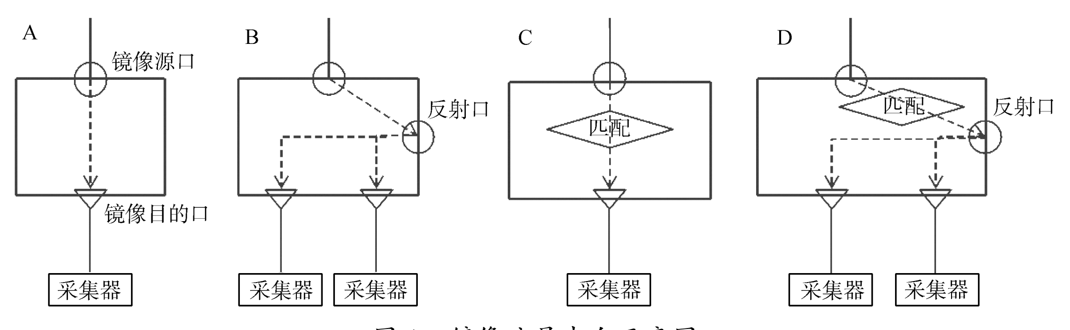


图 1　镜像流量走向示意图

本地镜像将数据在硬件转发时进行额外复制，发往镜像目的端口。其镜像数据流走向示意图如图 1 模型 A 所示。远程镜像技术将镜像报文发送至远程镜像反射口，通过镜像反射口将报文在远程镜像 VLAN 内通过广播方式复制发送，数据流走向如模型 B 所示。流镜像仅将匹配流分类条件的报文复制到指定目的地，如模型 C 所示。

本文提出一种技术设想：根据流镜像和多输出远程镜像的技术原理特征，如果能在交换机转发平台将流镜像的数据流直接发往远程镜像技术中的镜像反射口，再由反射口将报文广播发往所有配置为远程 VLAN 的端口，如图 1 中模型 D 所示，即将流镜像与远程镜像功能深度结合形成一种复合镜像技术，则该复合镜像技术可以在硬件层面整合实现可精确控制的多路输出，解决目前大多数网络设备上镜像过滤与多输出难以并存的矛盾，为数字资源利用分析系统提供数据过滤支持。

2.3　整体技术思路

信息采集分析系统一般分为数据采集过滤、数据存储管理、数据分析表示三个功能模块[9]，由采集系统本身的数据采集模块捕获全部数据后再进行数据过滤。

基于支持 IPv4/IPv6 双栈以及万兆网络环境数据过滤方法的数字资源利用分析系统在传统信息采集系

统架构的基础上，将数据过滤功能提前至数据采集前完成，整体设计为具有数据过滤、数据采集存储、数据分析表示三个功能模块，整体模型如图 2 所示：

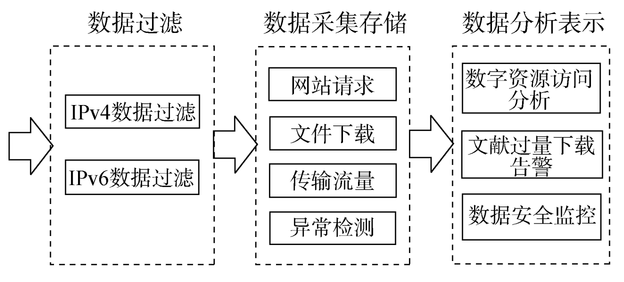


图 2　数字资源利用分析系统整体模型

数据过滤模块从核心网络获取一份同时包含 IPv4 及 IPv6 数据包的网络数据，通过复合镜像技术对数据进行过滤分流后仅将与图书馆数字资源相关的信息输出至可支持双栈流量的数据采集模块，以解决现有数字资源利用分析系统无法承载万兆双栈校园网网络流量的问题。

数据采集存储模块采集与数字资源相关的信息；数据分析表示模块提供数据展示、查询及报表功能，与传统系统功能基本相同。

3　实现方案

根据数字资源利用分析系统整体模型，系统共分为三个模块，其中数据采集存储及数据分析表示模块已有较多研究成果和成熟产品，本文研究重点为利用基于提出的复合镜像技术的数据过滤方法，在数据采集之前完成数据过滤。

3.1　具体技术问题

数据过滤模块的实现需解决以下技术问题：

(1) 镜像流量的再次镜像

作为一种技术设想，复合镜像技术不能直接应用于对稳定性有高度需求的生产网络，该模块需在独立于生产网络的网络硬件上部署完成。因此面临着该硬件能否识别从核心网络设备取得的数据源并进行转发的问题。

基于物理层的分路器传输以及基于数据链路层的端口镜像等输出技术都可用于从生产网络输出。鉴于分路传输在具有多链路负载均衡的校园网环境中部署不灵活且额外增加网络故障点的局限，本系统选用本地镜像方式。

网络设备通过匹配数据帧的目标 MAC 进行流量识别及转发，镜像技术作为网络设备管理功能，其报文镜像复制会在网络设备识别转发该报文前进行，其复制技术可以类比成 Hub[5]，即使网络设备因为 VLAN Tag 之类的传输标志不会转发某些镜像报文，也会先将报文先发往镜像目的端口，可以保证全部报文的镜像输出，实验同时证明了该机制。因此，基于输出后的数据报文通过独立的网络设备进行再次镜像输出是可行的。

(2) 复合镜像技术具体实现

交换机设备作为一个封闭的功能硬件，并不提供底层硬件转发的修改接口，为在实际环境中验证复合镜像技术的可行性，通过流镜像的流定义 ACL 对网络流量进行筛选过滤，仅将符合条件的流量送往某个端口，并对该端口进行再次远程镜像，实现可控的数据多路输出。该实验方案与复合镜像技术的差距在于用额外的物理接口转发了复合镜像技术设计中直接发往远程镜像反射口的流量。

实验证明直接对流镜像目的口进行远程镜像，远程端口目的端口并没有预期的输出。其原因为镜像目的口被标记为非转发口，不能直接对流镜像目的口进行再次远程镜像。为再次触发镜像功能，将流镜像目的端口与本机其他端口通过外部物理线路互联，将过滤后的流量重新送回网络输出系统，并对该回接端口进行远程镜像输出至多个端口，实现多输出。

(3) 对 IPv4/IPv6 双栈流量的支持

镜像过滤通过仅将匹配流分类条件的数据流复制至目的口实现。常规的流分类仅支持 IPv4 或者 IPv6 其中一种协议。支持双栈流量过滤需实现可以匹配 IPv6 流分类，并同时匹配 IPv4 流分类。

本系统已实现 IPv4/IPv6 流分类同时匹配。无法支持双协议同时匹配的硬件系统，也可通过对源数据分别进行复合镜像以及本地镜像加复合镜像并将数据输出至同一目的口的方式，实现对双栈流量的过滤。

3.2　部署方案

图 3 为数据过滤模块具体部署模型，主要基于一款中端的支持流镜像的数据中心三层万兆交换机实现。可见，校园网设备将一份镜像数据发往数据过滤模块所在的交换机，由该交换机将镜像数据进行 IPv4/IPv6 流量过滤后输出并回接至该交换机，并对回

接口进行多输出的远程镜像，发往数据采集硬件。数据过滤模块可为多个数据采集系统提供数据支撑。

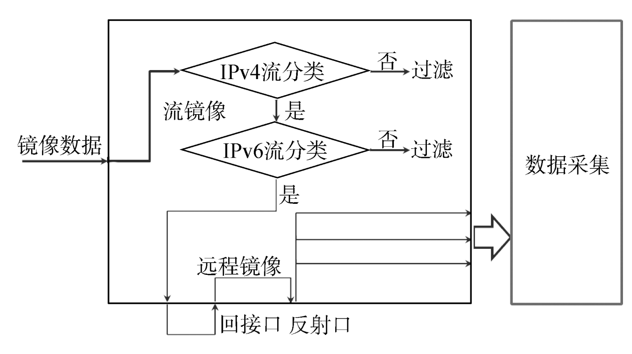


图 3 数据过滤模块部署模型

数据过滤模块的IPv4/IPv6流过滤规则如表1所示。IPv6流规则中的数据库资源组目前没有成员。

表 1 流过滤规则

行为	源 IP	源端口	目的 IP	目的端口	用途
deny	特殊用户	any	any	any	过滤特殊用户数据
deny	any	any	特殊用户	any	
permit	校图书馆	any	any	any	图书馆内网络用户通信数据
permit	any	any	校图书馆	any	
permit	数据库资源	any	any	any	校内用户对校外数字资源的访问
permit	any	any	数据库资源	any	
permit	图书馆资源	any	any	any	任意用户对校内图书馆资源的访问
permit	any	any	图书馆资源	any	
deny	any	any	any	any	过滤其他无关流量

数据采集与分析模块主要基于一款中端网络安全设备实现，主要实现的功能为数字资源访问分析、文件过量下载告警、数据安全监控等。

4 应用效果评估

4.1 数据过滤方法实验过程

4台三层交换机及两台PC终端被设计用于实现本文提出的数据过滤方法。Switch A/B/C用于模拟生产网络。Switch D为一款中端三层万兆交换机。终端A/B分别使用Wireshark抓取网络报文。

实验拓扑结构如图4所示，Switch B将G0/1端口的数据通过本地镜像发往G0/20口，G0/20口与Switch D的T1/0/4口互联。对T1/0/4口配置流镜像输出至T1/0/16口，通过光跳线将T1/0/16接回至T1/0/15，并对T1/0/15配置远程镜像输出至终端A/B进行报文抓取。

流控制信息配置如下：

```
acl number 3000
rule 0 permit ip destination 10.0.5.0 0.0.0.255
rule 5 permit ip destination 10.0.8.0 0.0.0.255
acl ipv6 number 3100
rule 10 permit ipv6 destination 2001: DA8: E800: 40: : /64
rule 15 permit ipv6 destination 2001: DA8: E800: 43: : /64
```

流分类配置如下：

```
traffic classifier mirror-2 operator or
 if-match acl 3000
 if-match acl ipv6 3100
```

Switch A分别往10.0.5.2等4个IPv4地址及2001: DA8: E800: 40: : 2等4个IPv6地址发5个Ping包验证。根据控制规则，终端A/B应同时捕获到发往10.0.5.2、

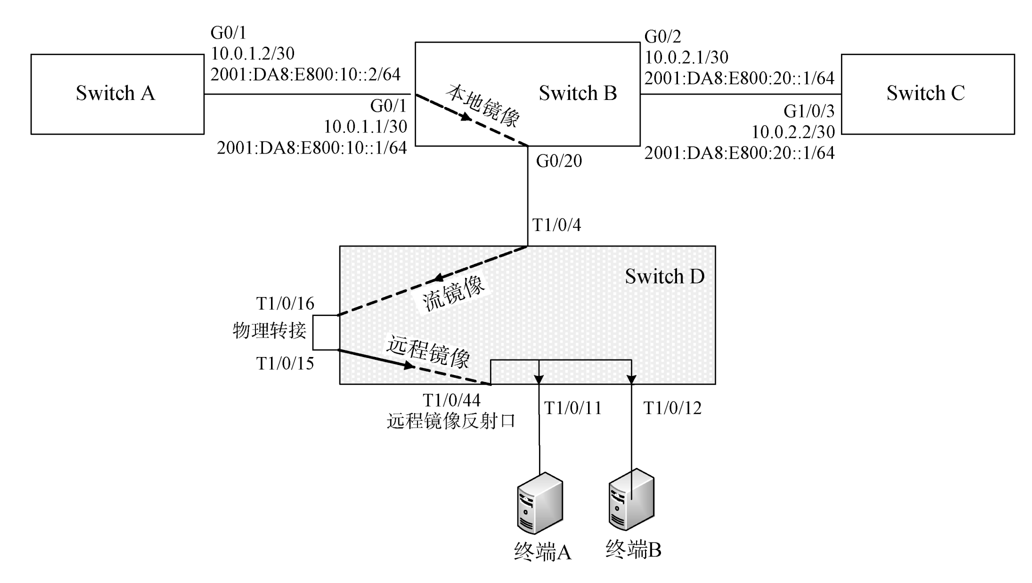


图 4 复合镜像功能实验拓扑

10.0.8.2、2001: DA8: E800: 40: : 2、2001: DA8: E800: 43: : 2的报文。其他报文被流规则过滤。

实际结果如表2所示。结果与预期相符，证明支持双栈的复合镜像技术可行，数据过滤方法准确有效。

表 2　实验预期及实际结果对比

目的 IP	预计报文数量	实际报文数量	
		终端 A	终端 B
10.0.5.2	5	5	5
10.0.6.2	0	0	0
10.0.7.2	0	0	0
10.0.8.2	5	5	5
2001: DA8: E800: 40: : 2	5	5	5
2001: DA8: E800: 41: : 2	0	0	0
2001: DA8: E800: 42: : 2	0	0	0
2001: DA8: E800: 43: : 2	5	5	5

4.2　系统运行情况

厦门大学图书馆已完成支持IPv4/IPv6双栈及万兆网络的数字资源利用分析系统的实验部署。

图5及图6为数据过滤模块过滤前后网络流量图，对比发现过滤后的数据量约为过滤前的1%-2%，常见的网络信息采集系统都可承载该级别的流量。

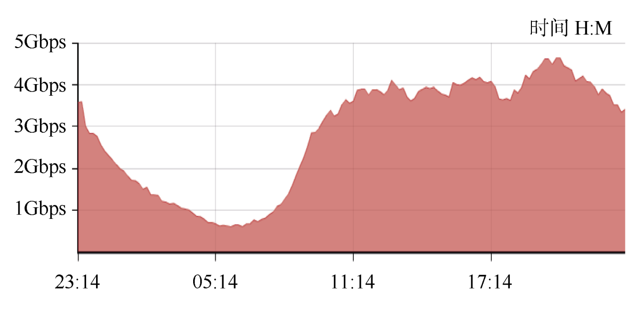


图 5　数据过滤前网络流量图

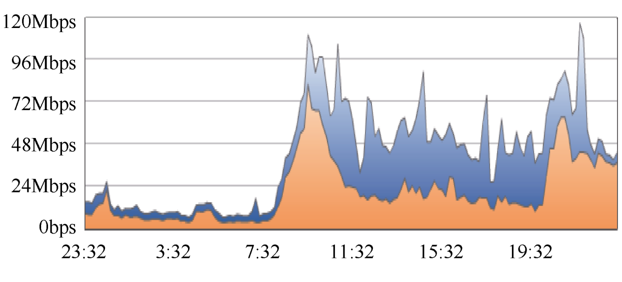


图 6　数据过滤后网络流量图

表3为数字资源利用分析系统与两组对比系统在同一天内采集的网页访问日志量。对比组为两款目前国内最高配置的安全管理设备，对比组1采集数据源为校园网IPv4/IPv6双栈出口镜像流量，即过滤前的网络数据；对比组2为校园网边界安全设备，不支持IPv6，采集数据源为校园网IPv4出口流量。

表 3　网页访问日志量对比

系统	网页	URL 关键字	日志条目数
数字资源采集系统	CNKI 概览页	epub.cnki.net/KNS/brief	18 390
	数据库介绍页	epub.cnki.net/KNS/html	691
对比组 1	CNKI 概览页	epub.cnki.net/KNS/brief	无法提供
	数据库介绍页	epub.cnki.net/KNS/html	684
对比组 2	CNKI 概览页	epub.cnki.net/KNS/brief	18 361
	数据库介绍页	epub.cnki.net/KNS/html	698

在承载大流量数据情况下，对比组1在日志功能及准确率方面不及本系统，对比组2与本系统所采集的数据结果接近。由此证明，本系统中的数据过滤模块可以稳定过滤无关流量并完整输出图书馆所关注的数据信息，数据采集存储模块足以承载过滤后的数据量。

数字资源利用分析系统主要功能为数字资源访问分析、文件过量下载告警、数据安全监控等。图7为以图书馆常见数据库为统计单位的站点请求量周报表；图8为数字资源常见数据格式文件下载量周报表；图9为安全防护功能中的病毒检测告警功能报表。

网站分类	访问用户数	站点数	总请求数
CNKI	2,861	86	144,687
ABI/INFORM	31	2	1,840
超星	102	5	1,657
Nature	49	2	401
ACS	65	5	398
CEIC	32	2	162

图 7　数据库访问量周报表

用户	传输工具	文件类型	总请求数
66.249.69.31	HTTP	pdf	10,327
66.249.69.15	HTTP	pdf	10,209
66.249.69.23	HTTP	pdf	9,944
66.249.75.224	HTTP	pdf	3,285
66.249.75.208	HTTP	pdf	3,273
66.249.75.216	HTTP	pdf	3,246
172.16.3.5	HTTP	caj	809
图书馆/4网段/210.34.4.79	HTTP	pdf	387
图书馆/4网段/210.34.4.79	HTTP	caj	187
图书馆/ireader/59.77.20.122	HTTP	caj	185

图 8　数字资源文件下载周报表

Top	病毒源IP	事件数	百分比
1	183.232.212.58	238	37.959%
2	2001:DA8:215:4131::17	34	5.423%
3	222.199.191.49	31	4.944%
4	115.156.188.148	24	3.828%
5	121.192.176.196	21	3.349%
其他		279	44.498%

图 9　病毒告警报表

4.3　系统缺陷

本文方法主要基于专业网络设备完成，需具备专业网络知识的人员进行配置维护，易用性较差；复合镜像技术的实现中采用端口回接方式，具有一定的部署风险。

5　结　语

本文提出复合镜像技术并成功验证其可行性，该技术可以解决当前网络环境中镜像技术多输出与流量过滤难以并存的问题；设计一种数据过滤方法，并成功部署支持 IPv4/IPv6 双栈及万兆校园网的数字资源利用分析系统，该系统部署成本低且具有较高的数据准确率。但该系统仅完成数字资源网络传输的记录及基础统计分析功能，提供的数据较为原始，未来应与数字资源利用评估体系进一步结合研究，以期为数字资源评估提供更为准确直观的数据依据。

参考文献：

[1] 潘补补．数字资源利用评估研究[J]．图书馆学研究，2012(13): 86-89. (Pan Bubu. Study on Utilization Assessment of Digital Resource [J]. Research on Library Science, 2012(13): 86-89.)

[2] 张静．图书馆用户数字资源利用行为实证研究[J]．广东工业大学学报：社会科学版，2012, 12(4): 74-78. (Zhang Jing. An Empirical Study of Library Users' Behavior in Utilizing Digital Resources[J]. Journal of Guangdong University of Technology: Social Sciences Edition, 2012, 12(4): 74-78.)

[3] Fusco F, Deri L. High Speed Network Traffic Analysis with Commodity Multi-core Systems[C]. In: Proceedings of the 10th Annual Conference on Internet Measurement. 2010.

[4] Bradner S. The Recommendation for the IP Next Generation Protocol [J/OL]. [2013-03-02]. https://www.rfc-editor.org/rfc/pdfrfc/rfc1752.txt.pdf.

[5] 马严．CNGI-CERNET2 成员单位积极部署 IPv6[J]．世界电信，2012(6): 61-62. (Ma Yan. Member of CNGI-CERENT2 Actively Deployed IPv6 [J]. World Telecommunications, 2012(6): 61-62.)

[6] 新一代互联网协议 IPv6 正式启动[J]．硅谷，2012(12): 14. (Next Generation Internet Protocol IPv6 Officially Launched [J]. Silicon Valley, 2012(12): 14.)

[7] Configuring Traffic Mirroring [EB/OL]. [2015-03-16]. http://www.cisco.com/c/en/us/td/docs/routers/asr9000/software/asr9k_r5-1/interfaces/configuration/guide/hc51xasr9kbook/hc51span.html.

[8] H3C．网络管理和监控配置[EB/OL].[2015-03-16]. http://download.h3c.com.cn/download.do?id=1034399. (H3C. Network Management and Monitoring Configuration Guide [EB/OL]. [2015-03-16]. http://download.h3c.com.cn/download.do?id=1034399.)

[9] 王继龙，吴建平．大规模计算机互联网络性能监控模型的设计与实现[J]．计算机研究与发展，2000, 37(4): 443-452. (Wang Jilong, Wu Jianping. An Internet Performance Monitoring Model Design and Implementation [J]. Journal of Computer Research and Development, 2000, 37(4): 443-452.)

作者贡献声明：

潘竹虹，萧德洪：提出研究思路，设计研究方案，论文起草及最终版本修订；

潘竹虹：采集、清洗和分析数据，进行实验。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

支撑数据[1-3]见期刊网络版 http://www.infotech.ac.cn；支撑数据[4]由作者自存储，E-mail：zhpan@xmu.edu.cn。

[1] 潘竹虹，萧德洪．细节模型.jpg．利用多种技术构建的数据过滤模块设计细节模型图.

[2] 潘竹虹，萧德洪．实验过程.rar．数据过滤模块各阶段实验拓扑及结果截图.

[3] 潘竹虹，萧德洪．病毒告警.jpg．数字资源利用分析系统病毒告警功能运行结果.

[4] 潘竹虹，萧德洪．现网配置.rar．厦门大学数字资源利用分析系统实际运行配置.

收稿日期: 2015-10-14
收修改稿日期: 2015-11-02

Data Filtering Method for Digital Resource Usage Analysis System for Dual Stack and High Speed Network

Pan Zhuhong[1] Xiao Dehong[1, 2]
[1](Information and Network Center, Xiamen University, Xiamen 361005, China)
[2](Library of Xiamen University, Xiamen 361005, China)

Abstract: [**Objective**] This study aims to promote the creation, management and use of academic library's digital resources. [**Context**] The development of IPv6 and 10 Gigabit Network generated difficulties in network data acquisition. [**Methods**] We proposed a port mirroring technology for network devices. Data from the IPv4 and IPv6 networks were filtered before they were collected for the digital resource usage analysis system. [**Results**] We built a practical digital resource usage analysis system. [**Conclusions**] The proposed method helped academic library establish digital resources analysis systems for the IPv4/IPv6 dual stack and high-speed campus network environment.

Keywords: Digital resource Information acquisition IPv6 Port mirroring

NISO 拟制定用于跟踪链接来源的推荐做法

美国国家标准组织(NISO)于近日批准了一个新的项目：制定互联信息环境下跟踪链接来源的推荐做法。一直以来，图书馆都在努力改进读者访问馆藏的方法，图书馆是否能够明确知道读者在获取到最终的图书馆馆藏内容之前是从如何开始他/她的检索之旅的，对于图书馆员在平台价值判断和资源分配上来说是非常重要的。此外，出版商也会通过 Web 日志分析来跟踪其用户的来源。在很多情况下，进行资源访问时使用一个关键的技术工具 —— 链接解析器 —— 会无意掩盖原本的引文来源。NISO 的这个新项目将为业界创造很多新的可能，允许内容主机和图书馆来确定链接的源头，这项新举措还能够提供有关如何获取和处理更多精确统计信息的指导。

"学术机构有为学生和教职员工提供方便、安全和有隐私控制的高质量受许可资源的使命，而图书馆和内容供应商又有收集有关这些资源被利用情况的需求，本项目将为这二者之间搭建起一座桥梁。"项目倡议者之一 EBSCO 信息服务集团高级副总裁 Scott Bernier 解释："虽然描述内容使用程度的统计数据一般来说都是可获取的，但是，本项目将提供一种方法，能够获取到这些使用如何发生。对每个访问请求有了精确的、持续的观察，图书馆就能拥有更多的决策支持信息，内容提供商也能获得更多有关其资源的曝光率及使用情况的信息。这些数据无疑会产生一定的影响，全面提升图书馆服务的价值。"

(编译自: http://www.niso.org/news/pr/view?item_key=c2ab810f6ee30db6113af9bd6638748d5caf94f7)

(本刊讯)